5J-06　XMLを用いたメーリングリストの知識抽出

瀬川 修　古池 一全　中部電力株式会社
杉江 修　村瀬 晋二　徐 一斌　株式会社シーティーアイ

1　はじめに

組織内での情報伝達や、グループワークの情報共有手段として登録メンバー間での同報メール(メーリングリスト)の利用が盛んに行なわれている。本研究ではメーリングリストのさらなる高度利用を目的として、XMLの情報構造化記述の枠組を応用した知識処理のためのマークアップ言語 **ML**^{2}(Mailing List Markup Language) を開発し、その処理系の初期実装を行なった。

2　メーリングリストの高度利用

従来のメーリングリストは投稿メールのIDによる管理や、ブラウザを利用した単純なスレッド表示の機能しかサポートされておらず、アーカイブに格納された情報の検索や再利用の面で必ずしも十分な機能が備わっているとはいえない。今後想定されるニーズとして、知識・ノウハウのデータベースとしての利用や、論旨追跡による高度な内容検索の支援機能などが考えられる。そこで、本研究では利用者がメール本文に明示的に意味的なマークアップを行なうことによって議論中に現れる質問-回答などの知識抽出を実現するシステムを開発した。

電子メールの知識処理に関する関連研究としては、本文の文書構造認識による情報抽出[1]や話題の連鎖による討議スレッドの自動抽出[2] などが行なわれているが、本研究ではXMLの共通フォーマットで構造化した電子メールのアーカイブから重要な知識や情報を精度よく抽出する手法を実現している点が特徴である。

3　ML^{2}を用いた知識抽出

3.1　概要

メール作成時に利用者が質問や回答、確認事項など内容に応じて本文に意味的なタグ(表1)を付与して投稿する。システムでは投稿されたメッセージごとにXMLファイルとして保存し、アーカイブされたファイルは **ML**^{2}の処理系によって質問-回答、確認-返事などの依存関係(通常のスレッドに対し**論理スレッド**と呼ぶ)が自動的に解析され、知識ベース化される。

表1: **ML**^{2}タグの例

タグ種類	依存関係
<質問>	<質問> ← <回答>
<確認>	<確認> ← <返事>
<回答>	—
<返事>	—
<情報提供>	—

3.2　処理系

システムでは前処理としてヘッダ情報のタグ付けを行ない、本文の意味タグの属性値として内部処理で用いる固有のID(ヘッダのメッセージID)を埋め込む。また本文中で引用のなされている箇所は自動的に<引用>タグを付与したXMLの構造に変換される。次にシステムではXMLファイルのアーカイブごとに一つのDOMツリーに展開してからタグの依存関係の解析を行ない、知識抽出した結果をDOMの推論木として出力する。

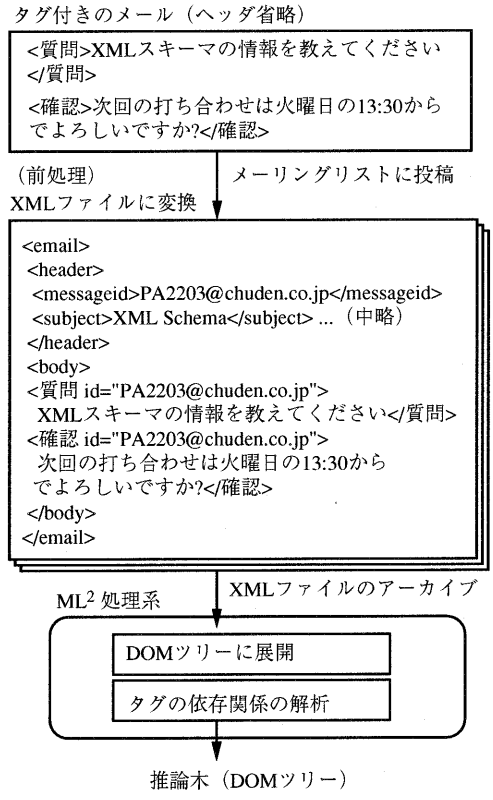

図 1: 処理の概要

Knowledge extraction from mailing-list archive using XML
Osamu Segawa, Kazumasa Koike (Chubu Electric Power)
Osamu Sugie, Shinji Murase, Yibin Xu (CTI)

4 知識抽出のアルゴリズム

前処理したXMLファイルのアーカイブに対し、

1. 物理スレッド(メッセージ単位のスレッド)の構造を保持して本文のタグを展開しDOMツリーに変換。
2. 展開したDOMツリー上で、
   i) <質問>、<確認>など依存関係が未解決のノードを探索する。ここでは<引用>ノード以下は対象にしない。
   ii) 未解決ノードが見つかれば、<質疑応答>などノード種類に応じた推論木(DOMツリー)のrootノードを新たに生成する。
3. 手順2で生成した推論木のrootノード全てについて、DOMツリー上で未解決ノードに対応するノードの探索を行なう。探索は以下の順で進める。
   i) <引用>ノードの子ノードに未解決ノードと同じノードがあれば(タグ属性値の固有IDとテキストを手がかりに判定)同じ<body>直下で<引用>ノードの直後→直前の順でペアの依存関係ノードを探す。
   ii) 上記(i)の探索で全ての依存関係が解決されない場合は、物理スレッド上で未解決ノードから最も距離の近い返信の<email>ノードの<body>直下でペアの依存関係ノードを探す。
4. 手順3で未解決ノードに対応するペアが見つかれば該当するノード以下の部分木を推論木に併合する。

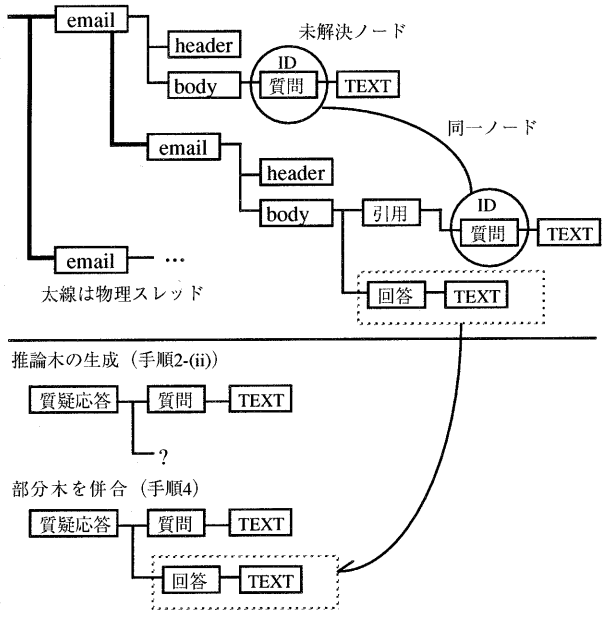

図2: 知識抽出のアルゴリズム

5 システムの実装

ML^{2}処理系の実装はJava言語によって行ない、パーザはXML Parser for Javaを用いた。

5.1 動作例

図3に知識抽出を行なった例を示す。結果はシステムで用意したXSLスタイルシートによってHTMLに変換することも可能である。

前処理後のXMLファイル(入力)

```
<in>
 <email>
  <header><messageid>PA2203@chuden.co.jp</messageid> ... (中略)</header>
  <body>
   <質問 id="PA2203@chuden.co.jp">XMLスキーマの情報を教えてください
   </質問>
   <確認 id="PA2203@chuden.co.jp">次回の打ち合わせは火曜日の13:30
    からでよろしいですか？</確認>
  </body>
 <email>
  <header><messageid>FA0268@chuden.co.jp</messageid> ... (中略)</header>
  <body>
   <引用>
    <質問 id="PA2203@chuden.co.jp">XMLスキーマの情報を教えて
    ください</質問>
   </引用>
   <回答 id="FA0268@chuden.co.jp">日本語の情報はあまりないです。
    W3Cのサイトを参照して下さい。</回答>
   <引用>
    <確認 id="PA2203@chuden.co.jp">次回の打ち合わせは火曜日の13:30
     からでよろしいですか?</確認>
   </引用>
   <返事 id="FA0268@chuden.co.jp">14:30からに変更していただけ
    ませんか?</返事>
  </body>
 </email>
 </email>
</in>
```

知識抽出結果(出力)

```
<out>
 <質疑応答>
  <質問 id="PA2203@chuden.co.jp">XMLスキーマの情報を教えてください
  </質問>
  <回答 id="FA0268@chuden.co.jp">日本語の情報はあまりないです。
   W3Cのサイトを参照して下さい。</回答>
 </質疑応答>
 <確認事項>
  <確認 id="PA2203@chuden.co.jp">次回の打ち合わせは火曜日の13:30
   からでよろしいですか?</確認>
  <返事 id="FA0268@chuden.co.jp">14:30からに変更していただけ
   ませんか?</返事>
 </確認事項>
</out>
```

図3: 知識抽出結果の一例

6 おわりに

メーリングリストの知識ベース化や議論の内容検索の支援機能として、XMLの共通フォーマットで構造化したメッセージのアーカイブから重要な知識や情報を精度よく抽出する方式を提案した。今後の課題としては、より複雑なディスカッションの論旨追跡などがあげられる。

参考文献

[1] 長谷川, 浅野, 堀井: "電子メールのインテリジェントサービス", 人工知能学会誌 Vol.14 No.6 pp.19-26 1999-11

[2] 山見, 村越, 島津, 落水: "電子メールを利用したコミュニケーションにおける討議スレッド自動抽出法の実装と評価", 情処研報SLP31 pp.31-44 2000-6